Javier Hernando

De: Mireia Farrús [mfarrus@gps.tsc.upc.edu]
Enviado el: martes, 11 de abril de 2006 15:22
Para: kirk.sullivan@ling.umu.se; erik.eriksson@ling.umu.se; javier@gps.tsc.upc.edu
Asunto: [Fwd: IAFL Barcelona 2006. Abstract Accepted as a POSTER]

Datos adjuntos: untitled-2.htm

untitled-2.htm

---------- Original Message ----------
Subject: IAFL Barcelona 2006. Abstract Accepted as a POSTER
From: "M. Teresa Turell" <teresa.turell@upf.edu>
Date: Mon, April 10, 2006 9:38 pm
To: mfarrus@gps.tsc.upc.edu

Dear Ms. Mireia Farrús:

I am writing on behalf of the IAFL Scientific Committee. I am delighted to inform you that your abstract "Dialect imitations in speaker recognition"
has been accepted for presentation as a POSTER at the forthcoming 2nd European IAFL Conference on ForensicLinguistics / Language and the Law, to be held from 14-16 September 2006 in Barcelona.

The next stage will be for you to register for the conference.
Registration forms, accommodation details and travel advice will be available from next week on the conference website at
http://www.iula.upf.edu/agenda/iafl_bcn_06

I'd be grateful if you would please try to register by the beginning of June if possible to help me with the planning of the programme. For those wishing to make travel arrangements, please note that the conference will run from early morning on 14th September to late afternoon on September 16th.

If you have any queries about the academic side of the conference, please contact M. Teresa Turell at teresa.turell@upf.edu. Any enquiries relating to registration should be directed to Marta Sánchez at iulaforensic@upf.edu.

I will be in touch as the weeks go by with further information as appropriate, and please check the website regularly for updates.

With good wishes. I look forward to seeing you in September in Barcelona!

M. Teresa Turell
On behalf of the Scientific Committee

Evaluator A: acceptable as a poster
Evaluator B: acceptable as a poster

Dialect Imitations in Speaker Recognition

Mireia Farrús[1], Erik Eriksson[2], Kirk P. H. Sullivan[2] and Javier Hernando[1]

[1]Universitat Politècnica de Catalunya, Catalonia

[2]Umeå University, Sweden

Abstract

Voice imitation and disguise are possible threats to the performance of a speaker recognition system and to the accuracy of earwitness descriptions. One common disguise is the modification of the own dialect or accent. In this paper, this kind of disguising is explored, using recordings from a well-known actor with considerable experience of dialect and accent imitation.

In order to see how successful his dialect imitations are and how the process of speaker discrimination is influenced by accent disguise, two sets human perception tests were constructed. One set focused on American and British English dialects, and one set on American and London English accents and Spanish-accented English. Each set consisted of three parts: a same-different speaker test, a same-different accent test, and a select the accent from a closed-set of options test. The results show that Johnny Depp is successful without his visual props and demonstrate a high correlation between the quality of the accent imitations and the failure of the human listeners to recognize that the voices come from the same speaker. The third parts of the experimental sets suggest the importance of familiarity with the accent that feeds into parts one and two. Spanish listeners, for example, are less accepting of the Spanish-accented English than non-Spanish speakers.

Finally, the same speech segments used in the perception test were used in an automatic speaker recognition experiment in order to compare the results and to check the robustness of the system in front of the voice changes. The results showed, once again, a low correlation between human and automatic speaker recognition.

Introduction

Dialect and accent with their specific phonetic feature sets are part of the high-level voice characteristics that can provide relevant information about a person's identity. Their importance to forensic linguistics has been pointed out by researchers such as Shuy (1995). Both dialect and accent are subject to imitation or alteration in order to

avoid or complicate the speaker's identification. Such changes to the voice could thus affect the outcome of criminal investigations. A detailed understanding of dialect and accent traits, and how they can be disguised is, therefore, essential to the pursuit of justice in criminal law.

This paper takes a two-pronged approach in its interrogation of dialect disguise and considers automatic speaker recognition and human speech perception. Such two-pronged approaches were used in the speaker (cf. dialect) imitation research conducted by Sullivan and Pelecanos (2001) and Zetterholm, Blomberg and Elenius (2004). The work by Sullivan and Pelecanos showed that the recognition system was capable of classifying the mimic attacks more appropriately than human listeners. The work by Zetterholm et al. found a minimal correlation between the speaker verification system they used and their human listeners in how they judged the 62 imitations in closeness to the target speaker.

The first-prong of the investigation presented in this paper, the human perception element, considers whether theatrical accent modification affects human perception in the process of voice discrimination. Stage dialects are a widely investigated phenomenon (e.g., Halloran 2003 and Machlin 1975) and are used in the theatre and film. These dialect imitations are not made to fool people, but rather to draw people into an imaginary world. A dialect imitation could however be used in a criminal setting and for this reason it is relevant to investigate how convincing dialects imitated for theatrical use are.

The second-prong investigates whether speaker recognition systems are vulnerable to dialect and accent disguise. Since dialectal imitations and disguises are based mainly on high-level features such as intonation and lexical characteristics, it is not expected *a priori* to find a low-level based automatic recognition system capable of classifying one speaker's talk according to the accent spoken. However, frequency characteristics can be affected by modification in intonation and acting. Stage dialect imitations provide a good way to interrogate possible vulnerabilities due to dialect and accent disguise in speaker recognition systems.

Method: Human Perception Experiment 1

Participants

Thirty-six males and twelve females aged between 14 and 54 years who were native

speakers of English or judged themselves to be advanced learners took part in the experiment. None of the participants reported a hearing problem. The participants were recruited by the experimental leaders and requests to friends to spread information about the experiment.

Material

From the following movies and extra material available on DVD, five three–four-second segments were selected: Speaker A: *Chocolat* (Irish), *Blow* (American), *Finding Neverland* (Scottish), *From Hell* (British English), *Secret Window* (American), and Speaker A's own voice (extra material), and Speaker B, another male American actor speaking General American.

Procedure

The participants undertook three tests in a web-browser environment. Before Test 1, demographic, hearing, and language competence data was collected. In Test 1 each participant was presented with 15 randomly constructed pairs of stimuli selected from the 25 available stimuli. They were given the instructions "You are going to hear a set of files. Each file contains TWO passages. Your task is to decide if they are passages spoken by the same speaker. Indicate your decision by selecting the circle Yes or No". In Test 2 each participant was presented with 15 randomly constructed pairs of stimuli selected from the 25 available stimuli. They were given the instructions "You are going to hear a set of files. Each file contains TWO passages. Your task is to decide if the passages are spoken by speakers of the same dialect/accent/regional variety or not. Indicate your decision by selecting the circle Yes or No". In Test 3 each participant was presented with 15 randomly selected stimuli from the 25 available stimuli. They were given the instructions, "After having listened to each file you are asked to choose from the drop-down list [American, English, Scottish, Irish, Welsh, Australian, New Zealand, South African, Spanish] which accent the speaker has."

Method: Human Perception Experiment 2

Participants

Ten males and seven females aged between 21 and 60 years who were native speakers of Spanish or judged themselves to be advanced learners took part in experiment. None of the participants reported a hearing problem. The participants were recruited by the

experimental leaders and requests to friends to spread information about the experiment.

Material

The Scottish and the Irish accents of Experiment 1 were exchanged for Speaker A's Spanish accented English (*Before Night Falls*) and Speaker C, a Spanish male actor speaking English with a strong Spanish Accent.

Procedure

The procedure was identical to Experiment 1.

Method: Automatic Speaker Recognition Experiment

Two evaluation conditions were designed. For both the speech files were parameterized using 20 filterbanks to form 20 MFCC for a frame size of 24 ms and a shift of 8 ms. Delta and acceleration coefficients were included. Speaker GMMs were used, formed from 32 Gaussian mixture components and trained with four of the five speech segments. Five tests were realized for each voice/dialect by alternating the training and test files: each time a different subset of four files was used for training with the remaining fifth file being used for testing.

In Evaluation Condition 1, Speaker A's natural voice, his imitated dialects and the voices of the Speakers B and C were used to train different speaker models. The same set of speakers and imitated dialects were used in the test phase. In Evaluation Condition 2, the set of speaker models was reduced to three: Speaker A's natural voice, Speaker B and Speaker C. Speaker A's imitated dialects were used for the Evaluation.

Results

The results of the human perception Tests 1 and 2 for Experiments 1 and 2 are shown in Table 1. The results of Test 3 for Experiments 1 and 2 are shown in Table 2. A_OAm indicates speaker A's own American accent, A_IEng indicates speaker A's Imitation of an English accent, A_ISc indicates speaker A's Imitation of a Scottish accent, A_IIr indicates his imitation of an Irish accent and A_ISpa his imitation of Spanish accent-English.

The results of the Automatic Speaker Recognition experiments are shown in Table 3 (Condition 1) and Table 4 (Condition 2).

	A_OAm		A_IEng		A_ISc		A_IIr		B	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
A_OAm	73	81								
A_IEng	13	*19*	84	90						
A_ISc	11	*6*	30	*22*	70	75				
A_IIr	17	*28*	53	*50*	26	*25*	73	75		
B	*32*	62	*6*	*13*	*3*	*6*	*5*	*16*	62	87

(a)

	A_OAm		A_IEng		A_ISpa		B		C	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
A_OAm	71	80								
A_IEng	35	*33*	76	67						
A_ISpa	0	*7*	12	*7*	82	87				
B	*35*	67	0	*53*	*6*	*13*	76	87		
C	*6*	*7*	*0*	*13*	*18*	40	*0*	*13*	59	67

(b)

Table 1. Percentage of Yes-responses in Experiments 1 (a) and 2 (b) for Test 1 (T1) (Speaker) and Test 2 (T2) (Dialect). Italics indicate when the correct answer was No.

	Response								
Stimulus	Am	Eng	Sc	Ir	Spa	SA	Aus	NZ	Welsh
A_OAm	*79*	2	0	2	1	1	9	4	1
A_IEng	3	*59*	5	3	1	3	10	14	1
A_ISc	0	7	*44*	37	1	3	2	2	3
A_IIr	18	19	10	*15*	0	5	10	6	17
B	*85*	8	2	0	0	2	2	0	0

(a)

	Response								
Stimulus	Am	Eng	Sc	Ir	Spa	SA	Aus	NZ	Welsh
A_OAm	*83*	3	0	0	0	0	10	5	0
A_IEng	23	*49*	9	6	0	3	6	3	3
A_ISpa	5	8	8	5	*49*	18	3	3	3
B	*83*	11	1	1	0	0	1	0	3
C	0	5	0	3	*63*	28	3	0	0

(b)

Table 2: Percentage dialect selection (test 3) by Experiment 1 (a) and 2 (b) listeners. Italics indicates correct dialect selection where correct is defined as the dialect Imitated for the A_I voices and the speakers actual dialect for the A_OAm, B and C voices.

	Assigned voice						
Test data	A_OAm	A_IEng	A_ISc	A_IIr	A_ISpa	B	C
A_OAm	4	0	0	0	1	0	0
A_IEng	0	5	0	0	0	0	0
A_ISc	0	0	1	2	0	1	1
A_IIr	0	0	0	5	0	0	0
A_ISpa	1	0	1	0	3	0	0
B	0	0	0	0	0	5	0
C	0	0	0	0	0	0	5

Table 3: Voice assignment from the Automatic Speech Recognition Condition 1 test where each of Speaker A's dialects formed a separate speaker model.

	Assigned voice		
Test data	A_OAm	B	C
A_IEng	0	1	4
A_ISc	0	1	4
A_IIr	0	0	5
A_ISpa	3	0	2

Table 4: Voice assignment from the Automatic Speech Recognition Condition 2 test where none of Speaker A's dialects formed a speaker model.

Discussion and Conclusion

The results of Experiments 1 and 2 show that Depp (Speaker A) was successful in his dialect imitations and that these accents would result in a witness describing his accent as one other than Depp's native. The results also suggest an interaction between accent and knowledge of the accent(s) in focus. This is seen in the difference between the Experiment 1 and 2 listeners' responses to the imitated English accent (Table 1), and the inability of the Experiment 1 listeners to deal with the imitated Irish accent (Tables 1a and 2). A similar, though much weaker, trend is seen in for the imitated Scottish voice. The difference could be due to Irish imitation being of poorer quality that the Scottish imitation or that more of the listeners were familiar with the Scottish accent. This can only be resolved by running the experiment groups of Scottish and Irish listeners.

Yet, in spite of the variation in the familiarity with the different accents, when same dialect and speaker segments were presented (T1), the yes response rate was

markedly stable. For Experiment 1, the strongest yes responses were for the imitated-English accent and for Experiment 2, the Spanish-accented English accent. However, as can be seen in Table 2, the listeners were less able to place these imitated accents; here the real American accents are the best placed. Of note is that Table 2b reveals that Depp has picked up on an aspect of the Spanish-accent English accent that led the respondents to select South Africa (SA) paralleling the responses to the native Spanish speaker.

The dialect task (T2), revealed that, with the exception of the responses to the imitated English accent by the Experiment 2 listeners, the recognition of the same dialect outscores recognition of the same speaker. Hence, it appears that accent-specific features dominate speaker-specific features (Eriksson, Schaefler, Sjöström, Sullivan & Zetterholm [forthcoming] for discussion).

The automatic speaker recognition experiment that used individual models for each of the imitated dialects (Evaluation Condition 1) resulted in little confusion between models. Further, only the imitated Scottish accent resulted in confusion between speakers A, B and C; Speakers B & C were perfectly recognized. However, under Evaluation Condition 2, where the three speaker models used were Speaker A's American Accent, Speaker B and Speaker C, and the imitated accents used as test data, only 3 of the 20 samples were correctly identified. These were imitated Spanish-accent English segments. This condition suggests that Depp's L1 dialect imitations include changes that affect automatic speaker recognition systems that Spanish accented-English imitation does not. In his L2 accented English he may have concentrated on other aspects of the voice, such as the feature that result in the selection of South Africa as the dialect (Human Perception Experiment 3; Table 3), that perhaps could never result in this imitated voice becoming distinct from his natural voice.

Together the experiments presented in this paper show that dialect imitation can confuse both the human listener and speaker recognition systems, yet in different ways, and that high-quality dialect disguise is a topic that warrants forensic linguistic consideration.

References

Eriksson, E.J.; Schaefler, F.; Sjöström, M.; Sullivan, K.P.H.; Zetterholm, E. (forthcoming*)* «On the perceptual dominance of dialect».

Halloran, N. (2003). The acquisition of a stage dialect. Department of Applied Linguisitcs, Portland State University, USA. [Master's thesis supervised by: Tucker Childs, PhD; Stephen Reder, PhD, and William Tate.]

Machlin, E. (1975). *Dialects for the stage.* New York: Routledge/Theater Arts.

Shuy, R.W. (1996). «Dialect as Evidence in Law Cases». *Journal of English Linguistics*, 23(1/2). 195-208.

Sullivan, K.P.H.; Pelecanos, J. (2001). «Revisiting Carl Bildt's impostor: Would a speaker verification system foil him?». In *Audio- and Video-based Biometric Person Authentication. Third International Conference, AVBPA 2001*, Halmstad, Sweden. 144-149.

Zetterholm, E.; Blomberg, M.; Elenius, D. (2004). «A comparison between human perception and a speaker verification system score of a voice imitation». In Proceedings of the 10th Australian International Conference on Speech Science & Technology (SST), Sydney, Australia. 393-397.

Acknowledgements

Mireia Farrús's participation was funded by grant AP2003-3598 from the Spanish Government.

Erik Eriksson and Kirk P H Sullivan's participation was funded by grant from the Bank of Sweden Tercentenary Foundation Dnr: K2002-1121:1-4 to Umeå University.